学校编码：10384　　分类号_____密级_____

学号：X2010230561　　UDC _____

厦门大学

硕 士 学 位 论 文

银行风险数据集市对公部分的设计与实现

Design and Implementation of Corporate Risking Data Mart in Banks

林丹旭

指导教师姓名：　赖永炫 助理教授

专 业 名 称：　软 件 工 程

论文提交日期：　2012 年 10 月

论文答辩日期：　2012 年 11 月

学位授予日期：　2012 年 11 月

答辩委员会主席：_____

评　　阅　　人：_____

2012 年 10 月

厦门大学学位论文原创性声明

本人呈交的学位论文是本人在导师指导下，独立完成的研究成果。本人在论文写作中参考其他个人或集体已经发表的研究成果，均在文中以适当方式明确标明，并符合法律规范和《厦门大学研究生学术活动规范（试行）》。

另外，该学位论文为（　　）课题（组）的研究成果，获得（　　）课题（组）经费或实验室的资助，在（　　）实验室完成。（请在以上括号内填写课题或课题组负责人或实验室名称，未有此项声明内容的，可以不作特别声明。）

声明人（签名）：

年　　月　　日

厦门大学学位论文著作权使用声明

本人同意厦门大学根据《中华人民共和国学位条例暂行实施办法》等规定保留和使用此学位论文，并向主管部门或其指定机构送交学位论文（包括纸质版和电子版），允许学位论文进入厦门大学图书馆及其数据库被查阅、借阅。本人同意厦门大学将学位论文加入全国博士、硕士学位论文共建单位数据库进行检索，将学位论文的标题和摘要汇编出版，采用影印、缩印或者其它方式合理复制学位论文。

本学位论文属于：

（　　）1.经厦门大学保密委员会审查核定的保密学位论文，于　　年　　月　　日解密，解密后适用上述授权。

（　　）2.不保密，适用上述授权。

（请在以上相应括号内打“√”或填上相应内容。保密学位论文应是已经厦门大学保密委员会审定过的学位论文，未经厦门大学保密委员会审定的学位论文均为公开学位论文。此声明栏不填写的，默认为公开学位论文，均适用上述授权。）

声明人（签名）：

年　　月　　日

摘 要

在国内投资热的环境影响下，随着银监会对银行业风险的管理和防范逐渐增强，银行业更加认识到管理风险的重要性，对风险管理工作越来越重视。但是由于银行数据的不完整性和业务系统的孤立性，银行无法建设全面的风险管理系统，如何从整体的角度管理风险已成为当务之急。

数据集市技术的发展，为风险管理提供了最佳的解决办法。数据集市能够为风险管理部门的信息需求设计数据模型、存储历史数据和加工业务指标。相较于数据仓库，数据集市面向特定主题组织数据的性质，虽然减少了一定的灵活性，但是增加了数据的使用效率。

风险数据集市建立在数据仓库之上，为风险管理系统提供面向主题的数据。由于整个风险数据集市的建设过于庞大，本文介绍的是风险数据集市对公部分。首先分析了项目研究的背景和意义，并介绍数据集市相关技术的发展现状，为风险数据集市对公部分确定研究方向和技术基础。同时，根据对风险管理业务的调研，归纳并分析集市的需求。经过总体设计和详细设计，确定了集市需要的运行环境，构建了风险数据集市对公部分的数据模型，设计了数据更新的 ETL 过程，并制定了数据质量检核的方法。最后通过确立系统配置和编写程序，实现了风险数据集市对公部分。

实践证明，风险数据集市对公部分的建设对于银行风险管理是具有实际意义的，可以对于银行风险管理建设提供借鉴。

关键词：数据集市；数据模型；ETL

Abstract

With the gradually evolutionary risk management and prevention of China Banking Regulatory Commission (CBRC) to the domestic banking, the banking industry starts to pay much more attention to the importance of risk management. However, many banks are not ale to build up a whole risk management mechanism due to the imperfections and isolation of the banking data. Therefore, how to manage the risk from a macro perspective becomes the preoccupation of the banking industry.

The development of data mart provides the risk management with the best resolution, because it can realize the needs of the sectors of risk management of designing data models, storing historical figures and processing operational indicators. Compared to the data warehouse, data mart focus mainly on the attributions of organizing data to certain targets. It makes the use of data more effective, even if it is less flexible than the data warehouse.

The risking data mart is based on the data warehouse and offers subject-oriented data to the risk management system. Given that the whole risking data mart is huge, this paper mainly discusses corporate risking data mart. First, it analyses the background and implications of this project. Second, it introduces the current development of the related techniques in the risking data mart, which aims to confine the topic in a research area. Meanwhile, according to the survey of the work in risk management, this paper analyses the needs of this mart. Third, through an overall design and detailed design, the thesis determines the working environment of the needs of the mart, builds up the data model of the corporate risking data mart, designs the ETL process of the data updates and works out a plan to check the data quality. Finally, it achieves the risking data mart via system configuration and programming.

The practice show that the build-up of corporate risking data mart is practical to the risk management of banks, which may be used as a reference to it.

Keywords: Data Mart；Data Model；ETL

目 录

CONTENTS

第一章 绪论

1.1 背景

随着加入 WTO 后，中国银行业对外开放程度的加大以及 2006 年底巴塞尔新资本协议在成员国间的实施，都无疑成为中国银行业在短期内面临的两次巨大外部挑战。商业银行作为经营货币业务的金融企业，风险管理是其经营管理的核心环节[1]。而较为落后的风险管理体系仍然是中国银行业，乃至整个中国金融体系稳定与繁荣的最大威胁，能否有效地对各种风险进行科学的管理和防范直接关系到银行自身的安全与发展，其中信用风险更显得重要[2]。

信用风险管理要是能够有效地被执行，除了制定适当的信用风险管理的政策与适时监督银行整体的风险外，更为积极的一种方法就是促使信用风险管理的理念深植于商业银行的组织文化中。同时要建立科学的信用风险管理体系[3]。首先要建立起全面的信用风险管理的模式。其次还要建立和完善信息系统及有效的交流渠道，使信息系统能够涵盖银行所有的业务活动，并且提供信用风险管理需要的指标数据。

鉴于银行业复杂的系统和海量的数据，数据口径维度众多，数据量及其庞大，采用数据仓库技术来实现决策分析系统是最佳的选择。数据集市是针对部门级应用的数据仓库，可以在数据仓库的基础上增加具有部门特色的功能需求[4]。因此，数据集市也就是银行风险管理系统的最佳选择。而数据仓库、数据集市、ETL 技术等相关技术的发展成熟为建立银行风险管理系统提供了强有力的技术支持。目前，大多数据银行都拥有自己的数据仓库或类数据仓库，服务于银行各类性的决策。同时为满足风险监管的需求，多数银行都在数据积累完成的基础上开发风险管理系统。运用数据集市实现风险管理系统来处理数据，展示报表成为一种趋势。本文将在这些基于数据集市的风险管理系统之前，建立一个统一的风险数据集市，满足各种风险管理系统对数据的需求，并提高效率和需求。

1.2 研究意义

本研究课题属于数据集市领域，开发基于数据仓库的银行风险数据集市。现在拥有数据仓库的大中型银行，其各风险管理系统都是直接使用数据仓库提供的数据。这种使用方式对于业务分析人员、项目开发人员和数据仓库都带来很大的负担，如下：

首先，数据仓库的面向主题性质使其数据被打散，对于不了解数据仓库的使用者难以理解和使用。各风险管理系统需要将分散的数据从新组织成易于理解使用的面向业务数据，并且检查数据质量。

其次，风险类数据由各个不同的源系统构成。每个风险管理系统都会使用到大量不同源系统的数据，使用者需要花大量精力理解数据含义、理清数据关系。再次，各个风险管理系统都有海量的明细数据需求，这些需求有许多都是共同的。每个系统都存储和加工可以通用的数据，既浪费有限的存储空间，也浪费加工处理的时间。

针对这些存在的问题，在风险管理系统和数据仓库中间建立一个风险数据集市，其功能是提供面向业务的数据，以及理解数据之间的逻辑，并且统一存储和管理。这样不仅节约的时间和空间，还使风险管理系统可以减少基础数据处理的投入，从而专注于处理业务逻辑。

1.3 本文的内容安排

本文主要介绍了风险数据集市系统中对公部分的设计与实现。研究的主要内容主要包括 3 个方面：

1、建立数据模型将在数据仓库分散在多个实体中的数据组合起来，提高数据对象的业务含义。

2、编写数据仓库数据到数据集市数据的 ETL 过程程序，以及验证数据集市数据正确性的数据检查程序。

3、构建 ETL 过程程序和数据检查程序在无人工参与下自动运行的调度。

全文一共分为 6 个章节，各章节介绍如下：

第一章绪论，主要介绍银行风险管理的背景和现状，并概要阐述了项目研究的目的和意义。

第二章数据集市相关技术，简要介绍了在集市分析设计中引用到的知识和技术，包括数据集市中技术介绍、ETL 概念描述和 TERADATA 关系型数据库说明。

第三章系统需求分析，通过建设目标的设置和业务需求的了解和分析，提出集市对公部分需要基本功能，以及对集市各种性能的要求。

第四章系统设计，在需求的基础上将集市系统按功能划分成数据获取、数据存储、数据检核、集市运行 4 个模块，描述了各模块与其他模块的联系。并通过细化讨论数据更新的 ETL 过程、数据模型的设计、数据检核机制，对于集市系统的进行了详细设计。

第五章系统实现，介绍系统的配置，以及集市数据存储、脚本编写和作业调度的过程，结合系统界面具体展现了集市实施的成果。

第六章总结与展望，对本文的研究内容和阶段性成果进行了总结，并对一下步工作做了展望。

第二章　数据集市相关技术介绍

2.1　数据集市介绍

2.1.1　数据集市的分类和由来

数据集市是面向部门级业务，面向某个特定的主题，是一种简化的小型数据仓库[5]。数据集市作为数据仓库中的一个技术，在其中起着极其重要的作用。由于种种原因，数据集市发展到如今，存在着两个发展方面。

1、独立的数据集市

一种数据集市是独立性的数据集市。企业在开始进行企业信息化时，为了节省企业的成本，并没有建立一个全局的数据仓库，而只是在各个部门建立局部的数据库，以后随着发展形成各个独立的数据集市。它们直接从各个应用系统中组织数据形成数据集市，即每一个数据集市中的数据可以来自于企业的不同部门的业务处理。集市中的数据是与主题相关的各部门数据的集成。这种数据集市虽然成本低，但存在着严重的弊端。由于各个数据集市的数据源是分开的，各个数据集市之间缺乏必要的联系，致使以后的各种查询、同步操作等功能有缺陷。
目前，大多数的企业都采用这种形式的数据集市。以下几点说明了这种建立成本低的数据集市的不足。

（1）数据的冗余。由于各个数据集市是相互隔离的，对于每个数据集市来说，必须要有一个整体数据的备份，这些数据中有不少通常是不必要的，这会加大企业的维护费用。

（2）流程的冗余。如果有数据仓库作为数据集市的数据源，数据仓库可以对各个数据集市的活动进行集中化；对于独立的数据集市来说，它们则必须把流程进行复制一份，这会增加企业的维护费用。

（3）非集成化。由于各个数据集市是由各个不同的团队所建立的，且各个数据集市是面向各个部门的，这种情况势必造成独立的数据集市无法集成化，不能兼顾其它部门的信息，无法形成一个真实反映企业全貌的视图[6]。